算法拒绝的三维动机理论

张语嫣[1] 许丽颖[2] 喻 丰[1] 丁晓军[3] 邬家骅[2] 赵 靓[4]

（[1]武汉大学哲学学院心理学系，武汉 430072）

（[2]清华大学社会科学学院心理学系，北京 100084）

（[3]西安交通大学人文社会科学学院哲学系，西安 710049）

（[4]武汉大学信息管理学院出版科学系，武汉 430072）

摘 要 算法拒绝意指尽管算法通常能比人类做出更准确的决策，但人们依然更偏好人类决策的现象。算法拒绝的三维动机理论归纳了算法主体怀疑、道德地位缺失和人类特性湮没这三个主要原因，分别对应信任、责任和掌控三种心理动机，并对应改变算法拒绝的三种可行方案：提高人类对算法的信任度，强化算法的主体责任，探索个性化算法设计以突显人对算法决策的控制力。未来研究可进一步以更社会性的视角探究算法拒绝的发生边界和其他可能动机。

关键词 算法决策，算法拒绝，心理动机，人-机器人交互

分类号 B849：C91

1 引言

决策在生活中普遍存在。传统上，决策过程的主体是人类自身（当事人或他人）。但由于人类决策能力存在局限（如受决策者经验和情绪影响），容易导致决策失误，造成不良决策后果。随着算法技术的发展与普及，为了克服人类决策的局限性，基于信息整合和理性计算原则的算法决策被广泛引入工作甚至日常生活中。算法决策(algorithmic decision making)是一系列相关概念的总称，包含增强决策、决策辅助、决策支持系统、专家系统、决策公式和计算机辅助等(Burton et al., 2020)，新近研究也将算法决策概念扩展于机器决策、机器人决策和人工智能决策等(Malle et al., 2015)。相较于人类决策，算法决策具有快速(Bonnefon et al., 2016)、准确(Donnelly, 2017; Lohr, 2016)、客观(Andrews et al., 2006)、普适(Esteva et al., 2017)和低耗(Common Cents Lab, 2017)等优点。凭借这些特点，算法决策在社会生活的各领

收稿日期：2021-07-08

* 国家社科基金青年项目（20CZX059）；国家自然科学基金青年项目（72101132）；中国博士后科学基金面上项目（2021M701960）

通信作者：许丽颖，E-mail: liyingxu@mail.tsinghua.edu.cn；
喻丰，E-mail: psychpedia@whu.edu.cn

域，包括医疗(Biró et al. 2021; van den Berg et al., 2017)、经济(Harvey et al., 2017; Kaya et al., 2017)、司法(Angwin et al., 2016)、交通(Badue et al., 2020; Fournier, 2016)、军事(Horowitz, 2016; Shin et al., 2019)以及日常生活(Roberts, 2017)中的应用都日益增多。

物理功能的算法技术早已将人类从许多费时费力的活动中解放出来(Parasuraman & Riley, 1997)；判断、规划和创造性思维等认知功能的算法决策技术也越来越普及(Chouard, 2016; Luo et al., 2019; Oliveira, 2017)。可以说，算法技术的进步给社会带来了革命性的变化。然而，算法决策的诸多优势似乎并没有让人们喜欢上它。哲学家 Bostrom(2014)认为机器代表人类做决定可能会导致灾难，Elon Musk 称自动机器的崛起是人类“最大的生存威胁”(McFarland, 2014)。普通民众也更偏好人类决策而非算法决策。研究表明，算法通常能比人类更准确地完成决策任务，然而，人们还是会选择人类决策或遵循人类建议做出决策。Dietvorst 等人(2015)将这种人们偏好人类决策而回避更准确的算法决策的现象称为算法拒绝(algorithm aversion)。算法拒绝并非只发生于人们了解到算法决策可能出错后，更多发生于人们并不了解算法表现之前，因此算法拒绝实际上就是对算法的偏见评估，即对算法的负面行为和态度(Jussupow et al., 2020)。

算法拒绝表现于知、情、行诸方面，即人们在认知上不认同算法决策，在情感上不喜欢算法决策，进而外显在行为上拒绝算法决策。首先是认知拒绝，主要体现在人们对算法能力的不信任(Prahl & Van Swol, 2017)。如尽管在司法体系中引入算法决策可以帮助对抗法律体系中人类固有的主观性和一些潜在错误(Andrew et al., 2009)，但在皮尤民意调查中，56%的受访者依然认为使用算法进行假释决定的风险评估是不可接受的(Smith, 2018)。在医疗上，病人也会更严厉地评判从算法而不是从同僚那里寻求建议的医生(Shaffer et al., 2013)。其次，情感拒绝体现在人们对算法决策的使用或其决策结果产生负面情绪(Lee, 2018; Leyer & Schneider, 2019)甚至道德责备(Voiklis et al., 2016)。如让人们评估完全相同的艺术作品，人们会更喜欢由人类而非算法创作的艺术作品(Jago, 2019)。当人类法官决策或算法司法决策出现错误时，人们对算法决策错误有更多的负面情感反应，也更可能在法庭失败时采取法外措施来干扰司法秩序(Ireland, 2020)。再次是行为拒绝，即相比人类决策，人们会更少的选择算法决策或对其利用程度更低，甚至以消极行为反抗算法决策的结果(Filiz et al., 2021)。如在股市决策实验中，人们更相信人类专家给出的相同预测，也更愿意接受人类而非算法建议(Önkal et al., 2009)。在医疗中，与人类提供的医疗服务相比，消费者更不愿意使用人工智能提供的医疗服务，并且愿意为此支付的费用也更少(Longoni et al., 2019)。

为什么人类会表现出在各个心理层面上对算法的拒绝？我们认为，面对算法决策这种新

的决策形式，人类大略会问这样三个递进的问题：第一，算法是否真的可以以及有能力进行决策？对这一问题的回答往往是负面的，人类通常情况下怀疑算法的能力，对算法产生怀疑和不信任，因此造成算法拒绝。第二，哪怕算法可以以及有能力进行决策，那么我们是否真的需要用算法进行决策，或者说算法决策对于人类个体有何益处吗？当然，对这一问题的回答通常也是负面的，因为人类在决策时有推脱责任的倾向，而算法作为决策主体，其道德主体地位和承担责任的能力缺失，导致人类采用算法决策无益，从而引起算法拒绝。第三，假使算法有能力且人类信任其进行决策，其也能进行道德责任的分担，这种决策模式对人类自身真的有良好影响吗？实际上对于这个问题的回答通常也是负面的，因为人类会因算法决策而缺失掌控，造成人类对于自己个性特征湮没而人将非人的感受，以此最终拒绝算法。为此我们提出算法拒绝的三维动机理论以说明算法拒绝的成因，并在此基础上探索其改变（见表1）：

表 1 算法拒绝的三维动机理论

算法决策的疑问	算法拒绝的原因	算法拒绝的动机	算法改变的方式
算法是否可以决策？	算法主体怀疑	信任/怀疑动机	提高对算法的信任度
是否需要用算法决策？	道德地位缺失	担责/推责动机	强化算法的主体责任
算法决策会对人造成何种影响？	人类特性湮没	掌控/失控动机	探索个性化算法设计

2 算法拒绝之原因

算法决策实际上是一个人机交互的过程，造成算法拒绝的原因也必不可少地出现在人、算法以及算法和人的交互中。但究其深层次原因，我们认为是三点，即算法主体怀疑、道德地位缺失以及人类特性湮没。这分别表现为人们倾向于认为算法作为新生事物而对其产生能力怀疑、人们倾向于认为算法无法作为道德主体而承担责任、以及人们倾向于认为算法普遍化的决策倾向减少了人类独特性。这回答了人类知觉算法决策时“能不能”（不了解算法决策也不觉得算法有能力决策）、“用不用”（算法决策之后无法归责也无用）以及“好不好”（算法决策反而消解个体作为人的独特性）的问题，也对应了信任/怀疑、担责/推责、掌控/失控这三种动机。

2.1 算法主体怀疑

算法决策能不能有效是第一个问题，实际上人们对这个问题的回答是倾向于否定的，表现出对算法的怀疑。这种算法主体怀疑的第一个原因可能是人们对算法并不熟悉。很少有技术在引入工作场所后立即被接受(Parasuraman & Riley, 1997)。一开始人们并不了解新技术，就很可能不喜欢或不信任一个新的自动化系统。但是随着算法决策的应用普及或人们与其接触日益增多，人们会对算法决策越来越了解，也会更接受算法决策。这种现象可用单纯曝光效应解释。一旦图像和声音等的外部刺激信息经常暴露在人的面前，人们对其喜爱程度就可能提高(Zajonc, 1968)。例如，大多数人已经习惯了来自气象模型而不是来自邻居的天气预报，因为气象模型几十年来一直被广泛使用；相反，关于时尚潮流的算法建议仍然相对较新，可能会面临更大的阻力(Logg et al., 2019)。日本人在日常生活中可能更熟悉机器人，因而更接受机器道德决策(Komatsu, 2016)。Ireland(2020)也认为，一般来说，随着人们越来越习惯于算法，在司法系统中，算法拒绝的现象可能会减弱或完全消失，算法可能会变得司空见惯，算法的错误会被视为人类法官犯下的错误。

两种因素影响人们对算法决策的熟悉度。其一是算法本身的透明度，其二是人们自身的专业度。首先，算法决策的不透明可能是阻碍人们熟悉算法决策的原因之一。算法决策是由黑箱方法产生的(Castelvecchi, 2016)，人们并不了解其内部工作方式(Kroll et al., 2017; 林少伟, 唐林垚, 2020)，统计预测背后的推理对于没有接受过统计学训练的人来说可能是不可及或神秘的(Önkal et al., 2009)。例如有研究发现，数学能力较差的人对算法决策的认同程度更低(Logg et al., 2019)。实际上，算法决策和人类决策都是不透明的，但人们以为可以通过自省(introspection)的方式了解其他人类决策的心理过程，这种耦合效应(Coupling effect)使人们对人类决策的不透明感知较低，却放大对算法决策的不透明感知。这种对透明度的感知不对称会削弱人们理解算法决策的信心，使人们在客观上本就不熟悉算法决策的情况下，在主观上也更低估自己对算法决策的理解，却因人类决策是非黑箱过程产生的错觉而高估自己对人类决策的理解，这种对算法决策和人类决策的透明度感知与主观理解不对称则会强化算法拒绝(Cadario et al., 2021)。因此关于算法是如何执行的解释既可以提高人们对算法决策的客观理解亦可以提高他们主观上对算法决策理解的信心，从而提高算法决策利用率 (Cadario et al., 2021; Yeomans et al., 2019)。其次，对应用领域的熟悉度与专业度也会影响人们对算法决策的接受度。研究发现，受过金融教育的人更善于处理财务信息，因此更容易接受算法建议(Lusardi & Mitchell, 2011)，而受教育程度较低的人则更少意识到其需要建议(Lee & Moray,

1992)。投资知识较少的人更拒绝算法帮助自己投资(Niszczota & Kaszás, 2020)。但也有研究发现，对自己能力过于自信的个体也可能排斥算法决策(Soll & Mannes, 2011)。例如，从一开始 Meehl(1954)提出算法在某些决策上表现得比人类更好时，就遭到了专家们的强烈质疑，相关领域的专家们不愿相信线性模型可以超越他们的判断，他们甚至对该结论持敌意态度。在预测任务中，专家没有意识到算法建议的价值，对算法提供给他们的信息充耳不闻并且坚持自己最初的判断，但不接受算法建议的专家的预测准确率甚至不如外行人(Logg et al., 2019)。这些研究结果提示，对应用领域的熟悉度和专业度与算法拒绝间可能存在着非线性关系，过弱或过强的熟悉度和专业度都会强化算法拒绝，未来的研究可通过比较熟悉度和专业度的不同水平，进一步考察二者间的关系。

算法主体怀疑的第二个原因可能是人们认为算法不具有良好的决策能力。人们通常对算法的专业能力表示怀疑，认为算法决策的表现不如人类决策(Dzindolet et al., 2002; Prahl & Van Swol, 2017)。例如，人们不愿遵循医疗算法建议的原因是他们不相信算法具备做出好建议的专业医学能力(Promberger & Baron, 2006)；即使是让人们评估完全一样的艺术作品，人们也更喜欢被告知是由人类而不是算法创造的作品(Jago, 2019)。这表明人们的确对算法决策存在一种其专业能力不如人类的偏见。即使在算法决策与人类决策准确度相近或犯同等程度错误时，人们依然更偏好人类决策(Gogoll & Uhl, 2018)，只有在算法具有相对于人类的明显优越表现时，人们才可能会选择算法决策(Bigman & Gray, 2018)。甚至有时人们明知算法决策的表现优于人类决策，仍不会选择算法决策(Grove & Lloyd, 2006)。专业能力欠缺还可以通过大量学习过程来进行弥补，但遗憾的是人们甚至倾向于认为算法决策不具备学习能力(Highhouse, 2008)。即人们通常会认为算法错误是系统性的，算法不能从它们的错误中学习并调整，但人类错误是随机的，人类能够从以往的错误中吸取教训并随着时间的推移而改进而机器不能。所以当算法产生错误时，人们可能会更倾向于绕过算法，不再使用算法决策(Filiz et al., 2021)。加之有些算法决策是不透明的，人们不清楚其来龙去脉(Angwin et al., 2016; O'Neil, 2017)，认为算法无法学习或调整的看法可能会更强烈。

2.2 道德地位缺失

算法拒绝的第二个原因是算法的道德地位缺失，因此算法无法承担决策后的责任。许多时候，人们遵从他人决策的目的是为了转移和削弱自己承担的责任，而如果算法缺失道德主体地位，则其在承担责任或者广义的决策上便显得无用，造成人们的算法拒绝。从实际表现

上来说，相对于非道德决策领域，一旦决策涉及道德领域，人们的算法拒绝倾向则更为明显。对于涉及人类生死问题的道德决策，即使机器决策带来了积极的结果，人们也宁愿选择一个普通水平的人类而不愿选择一个高水平的机器(Bigman et al., 2019)。同时，在经济博弈任务中，人们更喜欢将与他人报酬有关的决策任务委托给人类而非机器，因为人们认为与他人有关的任务涉及道德，而大多数人本能地不喜欢在道德领域使用机器(Gogoll & Uhl, 2018)。在投资方面，人们也更喜欢人类基金经理帮自己投资，尤其是当投资类型涉及到道德时，人们会表现出更强的算法拒绝，因为人们认为需要由道德能力(moral competence)水平更高的基金经理来判断投资一些在道德上有争议的公司是否是恰当的，而人们认为机器不具备这种能力(Niszczota & Kaszάs, 2020)。这可能是因为与其他决策不同，道德决策深深植根于情感之中，需要由具有完全道德地位的主体做出决策(Gray et al., 2017; Haidt, 2001)。而人们基于感知到的心智差异来判断谁拥有（或缺乏）完全的道德地位(Bastian et al., 2012; Gray et al., 2012)。心智是通过两个维度被感知的，即能动性(agency)和体验性(experience; Gray et al., 2007)。能动性是指思考、推理、计划和实现意图的能力；而体验性是指感受情绪和诸如疼痛等感觉的能力(Gray et al., 2012)。

算法通常被认为具有一定程度的道德主体地位，但这种道德主体地位感知又远弱于人类。比如，算法有一定的能动性(Gray & Wegner, 2012)，如它们可以进行复杂的计算；但完全意义的能动性不仅局限于原始的复杂计算，还应包括自我控制、计划、沟通和思考等能力(Gray et al., 2007)。也有其它研究认为，做出道德决策的主体应具有交互性（interactivity）、自主性（autonomy）和适应性(adaptability; Floridi & Sanders, 2004)，以及道德推理（moral reasoning）、自主行为（autonomous action）、沟通和判断行为后果(Cushman, 2008)等能力(Malle, 2016; Malle & Scheutz, 2014)。而机器不具备这些能力，因此人们也认为机器不具备做出道德决策的能力。除了能动性以外，体验性对于道德决策也很重要。情绪对道德决策至关重要(Greene et al., 2001; Haidt, 2001; Haidt et al., 1993; Koenigs et al., 2007)，尤其是移情的能力，即感受他人痛苦的能力，似乎是道德判断的核心要素(Aaltola, 2014; Decety & Cowell, 2014)。而机器似乎缺乏感受真实情感的能力(Reinecke et al., 2021)。因此，尽管机器具有一定程度的能动性，但它们缺乏体验性(Brink et al., 2019)，缺乏感受道德情绪的能力(Malle & Scheutz, 2014)，它们依然不具备道德决策能力。

最为重要的是，若算法的道德主体地位不足，其便无法很好地承担道德责任。一旦算法无法承担道德责任，则人类在决策中便缺乏将责任转移给算法的可能，从道德责任分担意义上来说，算法决策便无可取之处(Bonaccio & Dalal, 2006)。换句话说，当遵循或纳入人类决

策者的建议时，人们会感觉自己对该决策的责任转移到了建议者身上，但若遵循或纳入算法决策，这种责任则不会转移，因为人们认为算法没有承担责任的能力(Bonaccio & Dalal, 2006)。Armstrong(1980)发现，即使有压倒性的证据表明专家的判断和建议有时并不比普通人更准确，人们仍然更相信这些专家。这种现象在政治预测、冲突结果预测和股市预测等不同领域都有发现(Green & Armstrong, 2007; Tetlock, 2009)。Armstrong(1980)认为，这种对专家依赖的一个原因是责任转移，即如果预测结果不准确，专家会受到更多的责备，人们就可以转移自己决策失误的责任。而 Bonaccio 和 Dalal(2006)认为，只有当决策建议者是人类的时候，分担和推卸责任这样的动机才起作用，若建议来自算法，这种动机便不起作用了。因为人类被认为有能力承担责任，而算法则没有(Promberger & Baron, 2006)。例如，在医疗上，当患者不得不决定重要的医疗程序时，他们不愿承担该决策的责任，而是将其转移给其他人——人类医生，因为遵循医生的建议可以让病人感觉自己不需要承担太多责任，但是遵循算法建议不会以同样的方式减少他们的责任感知(Promberger & Baron, 2006)。

2.3 人类特性湮没

人类特性湮没是算法拒绝的第三个原因，即人类感知到算法决策对人类身份或独特性的更具象征性的威胁，因无法展示出自身的独特性而失去控制感。首先，从人类个体来看，通常情况下，人们倾向于认为自己是独特而不同于他人的，这种独特性寻求在个体主义文化下尤甚(Brewer, 1991)。而算法相对理性，在没有个人数据的情况下只能以标准化和模式化的方式操作，以同样的方式处理每一种情况(Haslam, 2006)，这两种基本信念的不匹配使人们抵触算法决策。研究发现，与人类医疗服务者相比，消费者更不愿意使用人工智能医疗服务，因为人工智能医疗服务引发了人们的一种担忧，即一个人的独特特征、环境和症状将被忽视，这种担忧被称为独特性忽视(Longoni et al., 2019)。即例如，在消费领域，产品特征和消费者需求的信息不对称会伤害企业与消费者之间的关系(Van Swol, 2009)。人们在更主观的任务中更不愿意使用算法(Castelo et al., 2019)，如人们更依赖朋友而不是算法推荐电影与书籍或讲笑话，因为这些决策任务受个人品味支配(Yeomans et al., 2019)；相反，在体育预测等具有具体外部准确性标准的领域，人们可能会对算法建议感到更信任(Logg et al., 2019)。

算法决策带来的个性消失更深层次的原因可能是对整个人类群体身份与独特性湮没的担忧，即人们感知到算法决策对人类内部群体的独特性和价值观的威胁，甚至感觉自己被非人化（dehumanization）。人类认为自己与其它群体截然不同，但算法决策尤其是高自主性的

算法决策对环境的掌控可能会模糊“人类”与“工具”之间的界限。例如人们更讨厌不听人类指令而能自主根据环境而做出判断与决策的机器人，感知到的威胁也更高(Złotowski et al., 2017)。申请研究生时仅通过数据信息而不是面对面的交谈也被认为是不人道的(Dawes, 1979)。因此，提高主体参与或许能够提高算法认同。相比单独使用算法决策，人们更偏好算法与专家的共同决策，只要算法辅助的使用不取代人类的判断；对于不使用算法辅助决策的专家和使用算法辅助共同做决策的专家，人们甚至更倾向于后者(Palmeira & Spassova, 2015)。Pezzo 等人(2006)的研究也表明，与没有算法支持的医生相比，使用算法支持的医生犯错时，人们对其负面评价更小。即使是一个不完美的专家加入决策过程中，也可能会更轻松地排除障碍，做出更好的决策(Kuncel, 2008)。当然，研究也发现，如果算法决策不是替代而只是辅助，最终决策还是由人来决定，人们则不会感受到人类特性的湮灭，此时算法拒绝便会减弱。例如，相对于执行算法，人们更容易接受和喜欢咨询算法，也就是说，相比算法代替自己决策，人们更能接受算法给出决策建议(Dietvorst et al., 2015)。将机器限定在从属于人类的角色，而由人类做出最终决定，可以在一定程度上减轻人们对算法决策的拒绝(Bigman & Gray, 2018)。Longoni 等人(2019)也发现，当医疗人工智能只是辅助而不是替代人类医生做决策时，并没有出现对医疗人工智能的抵制。若人们能自主修改算法决策结果以掌握最终决策权，人们也会更接受算法决策(Dietvorst et al., 2018)。

3 算法拒绝之改变

算法和算法决策在人工智能时代不可避免地渗入人类生活，且实际上已有大量算法在驱动人类的信息获取及理解（如网站推荐算法），只是人们尚未意识到。算法决策不可避免地在未来会获得越来越多的应用。虽算法决策还在迅速发展的阶段，在一些领域还不够成熟，但其在许多问题上的决策能力已然超越人类，而人类出于种种原因尚未接受它，若能使人们欣赏其效用，提高对算法决策的认同与接受度，或许能为许多个人实际问题的解决带来事半功倍的效果，同时为社会带来经济效益。因此，针对算法主体怀疑、道德地位缺失以及人类特性湮没三种算法拒绝的原因——实际上，这三种原因对应着人类的三种心理动机，即信任、责任和掌控，我们提出提高算法认同的三种方式：提高人类对算法的信任度、强化算法的主体责任，以及探索个性化算法设计以突显人对算法决策的控制力。

3.1 提高对算法的信任度

对抗算法主体怀疑的核心因素可能是信任。不难发现，信任贯穿于算法拒绝的大部分原因之中。人类自身的认知特征如在算法方面的经验不足、对算法能力的感知不足，以及在算法决策中的参与度不足都可能导致对算法决策的不信任或信任脆弱，从而拒绝算法决策。对算法决策的不信任体现在两方面。第一，人们对算法决策的信任本就不如人类决策。相比人类决策，人们在了解到算法决策不完美或经历其失误之前，就已经不那么信任算法决策(Longoni, et al., 2019)。第二，人们对算法决策的信任更脆弱，这意味着，与人类相比，一两次算法方面的糟糕经历对其信任造成的伤害更大(Prahl & Van Swol, 2017)。因此人们能容忍人类的失误，却不接受算法的失误。Dietvorst 等人(2018)就发现，人们对算法决策错误的容忍度更低，当他们看见算法决策错误时，对其信心会陡然下降，之后的选择也会避开算法决策。因此，人们拒绝算法决策的很大一部分原因是人们没有建立起对算法决策的信任(Lee & Seppelt, 2006)，或者说算法没有“说服”人类相信其决策能力。

就算法拒绝研究来说，提高信任以减少对算法主体及算法能力的怀疑，最简单的办法当然是提高人类对算法的专业知识和熟悉性，并同时更多地展示算法能力。随着越来越多算法和人工智能科技元素在人类生活中的出现，实际上人类对算法决策的知识和熟悉性会逐渐提高直至习以为常。这时，展示算法能力便尤为重要。比如，专业和有效的算法建议会使人们对其更信任从而提高算法利用度(Goodyear, 2016; Kramer et al., 2018)；人工智能医疗服务能给医疗领域带来革命性变化的前提就是其可以达到专家级的准确性(Leachman & Merlino, 2017)。此外，若人们能根据自己的需求自主调整算法输出(Greene et al., 2016)，他们对算法决策学习或调整能力的感知也会更强，虽然算法决策还远非完美，但人们实际上愿意接受虽会犯错但可以学习或调整的算法(Berger et al., 2021)。不过，现阶段算法决策的专业性似乎只有在与人类决策进行比较并明显优于人类时才能得到凸显(Bigman & Gray, 2018)。

3.2 强化算法的主体责任

人们偏好人类决策而不喜欢算法决策可能出于分担和推卸责任的动机。对于人类来说，若出于类似动机则必然需要算法承担道德责任，这也意味着需要强化算法作为道德主体的道德地位，即让算法更像人。这种所谓“像人”有两种方式，一种是让其看起来像人，而另一种是让其似乎具有人类的定义性能力。先说后一种，即提高人们对算法心智能力的感知实际上可以提高大众对算法决策的接受度。当算法或人工智能看起来越具有类似于人类的心智能

力，人们就越相信它能完全胜任其预期功能(Waytz et al., 2014)。因为有意识的行为主体会被认为更能控制自己的行为，因而更能通过有意识的预测和计划来成功完成任务并对其行为与结果负责(Cushman, 2008)。例如，提高算法的感知情感相似性可以有效地增加算法在主观任务中的使用(Castelo et al., 2019)。

而对于以人工智能体如机器、机器人和自动驾驶汽车等为载体的算法决策，则可以通过拟人化过程提高其类人程度以强化其道德主体地位。拟人化是指将人类独有的本质特征赋予非人对象的心理过程(Waytz et al., 2010; 许丽颖 等, 2017)，例如可以从物理特征（脸部、眼睛、身体、动作）、心理特征（喜好，幽默，个性，感情，同理心）、语言（口语、语言识别）、社会动态（合作，鼓励，回答问题，互惠）和社会角色（医生、队友、对手、老师、宠物、导游）等方面增加算法或人工智能体的类人线索(Fogg, 2002)。重要的是，籍由拟人化还能影响心智感知尤其是理性思维（能动性）和意识感觉(体验性; Gray et al., 2017)。研究发现，在心智知觉上将人工智能体拟人化，确实可以提高人们对它们的信任，以自动驾驶汽车为例，当其能感知和思考周围环境，而不仅仅是一个没有脑子的机器时，它们会看起来更善于在车流中穿梭驾驶(Waytz et al., 2014)。此外，对人工智能体外在特征的拟人化可以间接提高人们对其心智能力的感知，进而更信任该人工智能体。例如，当自动驾驶汽车的外在物理特征被拟人化时，如赋予其名字、性别和类似于人类的声音，人们会更信任它，如果发生了由他人错误造成的意外事故，人们对拟人化自动驾驶汽车的责备会更小；相反，若自动驾驶汽车成功规避了该意外事故，人们会将成功更多地归因于拟人化自动驾驶汽车(Waytz et al., 2014)。当然，过度的拟人化或心智能力感知也可能适得其反，当算法或人工智能体在外观或心智能力上与人类达到一定相似度时，人们对其好感会陡然下降，甚至感到不安与害怕(Gray & Wegner, 2012)，因而拒绝算法决策，此即恐怖谷效应(uncanny valley effect; Mori, 1970)。因此，当赋予算法或人工智能体以人类独特特征时，应当控制在让人们感到舒适的范围内，从而使人们更认同与信任算法决策。

必须说明的是，尽管研究已发现算法在预测任务(Jordan & Mitchell, 2015)、棋类游戏(Dockrill, 2017; Silver, 2017)以及股市回报最大化(Zuckerman, 2019)等各种决策任务中表现优于人类，但道德相对而言更主观(喻丰, 韩婷婷, 2018)，就像在道德困境问题上，有人秉持义务论而有人秉持功利主义，暂无明确证据表明算法是否能比人类做出更好的道德决策。算法认同是指能欣赏算法决策的效用，对算法的积极态度和行为(Jussupow et al., 2020)，而不是对算法决策的盲目接受，尤其在面临道德问题时，应更审慎地看待算法决策。哪怕改进算法让人们产生了算法拟人化过程，问题本质并没有改变，即或许可以提高人们对算法承担责任

的能力感知以减轻算法拒绝，但很有可能并不能完全打消人们对责任归因的顾虑，因此更重要的问题是研究算法或人工智能的责任归因问题，到底算法决策错误应由谁负责，设计者、生产者亦或是使用者？随着算法决策日益普及，这是一个在未来必须面对的问题，也是减少或消除人们对于算法决策责任归因顾虑的最有效方法。哪怕算法能够最终进行责任分担，我们如何对其进行惩罚也将是面临的一个难题。在这一点上，也许同时对人类进行道德规范教育，让其更多承担道德责任并更少推责也是并行不悖的方案。

3.3 探索个性化算法设计

人类特性湮没的表象是无法展示出自身的独特性，而背后潜藏的动机即人类缺乏对于外部世界的掌控，或者说算法决策让人类感知到自主性(autonomy)丧失。为了还原人类的控制感，让人感知到自己作为独特个体的存在，探索个性化算法设计以突显人类对算法决策的控制力可能是较好的选择。这可能体现在满足人类个性化需求上，也可能表现在人类对于决策的最终裁定上。

首先，当人们认为算法可以了解自己的个人偏好时，算法决策会变得更容易被认同与欣赏。例如，在消费领域，能预测消费者对商品感知吸引力的电子商品筛选工具，可以从海量商品中优先向消费者推荐其更喜欢、更有可能购买的产品(Diehl et al., 2003)，既能减少顾客搜索优质产品的时间成本，也能提高商家成功卖出产品的可能性(Senecal & Nantel, 2004)。在医疗领域，对独特性的关注——即医疗人工智能若能提供更个性化与更具有针对性的服务，能缓解人们对医疗人工智能的抵制(Longoni et al., 2019)。在投资领域，更个性化的投资建议也会更受消费者的喜爱与信任(Alserda et al., 2019; Lourenço et al., 2020)。

虽个性化算法决策已然是发展趋势，比如许多公司使用算法推荐新闻或其他信息，但这也会造成新的问题。第一是回声室效应(Echo Chamber Effect; Cinelli et al., 2021)。现在的信息传播媒介大量采用算法，其基本原理是个体对于某种事物的偏爱会增加推荐频率，而这种所谓的偏爱实际上可能是基于其历史信息获取的经验共现，也可能是基于简单的标签。无论是何种方式，这种信息回声室效应也为人编织了一个信息的房间，它屏蔽其他信息，让个体在某一类信息中娱乐至死，这便是信息茧房。第二便是隐私问题。个性化算法决策离不开对个人数据的收集。实际上人们并不希望在所有领域都提供个性化算法服务。研究发现，人们对商业应用（如购物和娱乐）方面的个性化服务态度更积极，但反对新闻来源、社交媒体、政治竞选的个性化，反对算法收集和使用敏感的个人信息(Ipsos Mori, 2020; Pew Research

Center, 2019)。人们需要的不仅是个性化算法，更是透明的个性化算法，即尊重人们的数据隐私，并且可以受用户调整(Kozyreva et al., 2021)。为了提高人类个性化感知而过度收集数据，侵犯人们的隐私或在人们敏感的领域越界应用，可能适得其反。

其次，应让人类在决策系统中拥有最终发言权。未来我们将看到越来越多的机器在人类共同的环境中生活，而让算法完全替代人类决策或取代人类的工作从来不是一个好的选择。人类和算法擅长的方面是不一样的，算法或许比我们更能把一件事情“做好”，但只有人类才知道要“做什么”和“怎么做”。利用算法决策强大的计算与整合信息能力的优点，来弥补人类决策的缺点，将人类的智慧运用在人类更擅长而算法不可及的地方，相辅相成，以算法辅助人类或人机合作决策，才是更好的选择。因此，加强人类和算法的合作决策，同时确保人类在决策中的主体地位，以算法辅助人类而不是替代人类能使人们对算法决策持更开放的态度。总之，相较于单独的算法决策或人类决策，人们可能认为人类与算法合作能做出更优的决策(Palmeira & Spassova, 2015; M. V. Pezzo & S. P. Pezzo, 2006)，但前提是人类拥有最终决策权(Starke & Lünich, 2020)。我国学者钱学森早在上个世纪就已提出综合集成研讨厅方法，提倡以人为主，人机结合，不要用计算机代替人，而是要用计算机去协助人，将计算机高速处理信息的能力和人的综合思维能力（包括逻辑思维、形象思维和创造思维）结合起来(黄志澄, 2005)。钱学森的想法是基于人类至上的角度考虑问题，将人类的智慧和利益放在首位(李月白, 江晓原, 2019)，这对于今天的算法决策研究仍有深刻启发。

4 讨论与总结

算法拒绝的三维动机理论实际上是在模拟人类面临算法决策时的直觉思维框架，也就是人类在面临算法决策时，大略会问的几个递进问题，即是否了解算法并认为其有能力进行决策、是否使用算法而导致决策失败后自己承担责任以及是否使用算法会显得自己缺乏作为人类个体的独特性。这三个问题彰示出人类信任/怀疑、担责/推责、掌控/失控的三种动机。虽已综述现有研究并尝试理论建构，但仍须承认尚有诸多待探讨空间。

其一，此理论似乎暗含一种算法决策强于人类决策，而将算法决策的接受视为“应该”的倾向。实际上，人类并非因理性而决策，甚至人之所以为人正是因为人有独特于机器的理性之处。研究发现，在人工智能时代人类会将那些算法比人强的能力（如一般认知能力、身体能力、消极情绪等）视作不重要，而将道德、审美等能力凸显以区分自己与机器(喻丰, 2020)。人性光辉有时就闪现于情感而非机器理性中。同时，人类接受或者不接受某种事物，并非一

定源于其相较于人类具备优势，人类不会因为某种制度有优势就理应接受之，也不会因为某种商品更物美价廉便理应抛弃其他商品。不以事实描述的优劣做规范应该的判断依据本身也是人类特征，否则人便于算法无异。实际上该理论并未预设这种规范性立场，只是在探究为何算法在某些方面强于人类，而人类依旧选择拒绝算法的原因。当然，此理论确有期望人类可以利用算法决策之优势来优化自身决策的可能，但并非在强调必须用算法决策替代人类决策。

其二，算法拒绝的三维动机理论是开放的，并不拒绝其他可能动机。例如，普遍意义上更广泛的认知、存在以及社会动机均有可能作为备择动机(Jost et al., 2003; Jost et al., 2009)。当然，我们所描述的信任、责任和掌控本身就从属于广义的社会或者认知动机，但是确实存在其他可能。如认知闭合需求，此动机包含对明确且完美答案的渴望，因此高认知闭合需求的个体思想封闭，对开放性和不确定感到不安，进而对悬而未决与不确定的决策感到烦闷与厌恶(Otto et al., 2016)。算法决策作为现代信息时代的产物，其新颖性对于高认知闭合需求的个体来说，则可能意味着开放与不确定，从而引起算法拒绝。以往研究的确发现，算法拒绝很可能出现在人们看到算法出错之后(Dietvorst et al., 2020)，这可能是对完美答案渴望的认知闭合需要与发现算法失误的期望偏差之间的冲突导致的算法拒绝。再如社会认同，在传统的群际理论中(Turner et al., 1987)，最广泛的社会认同即人类认同，其反映了自我作为人类的认知，以及与非人类（即动物、其他生物、非生物实体）的潜在不同。但近年来研究发现社会认同在向非人实体扩张，如人与动物之间可能存在心理联结，这种联结包括三个维度，即团结、自豪与感知相似性(Amiot et al., 2020)。这种心理联结显然会使人类对动物产生更积极的情感联结而更喜欢动物。算法与动物都属于非人实体，若以动物类比，对算法缺乏心理联结或许会导致对算法的不喜爱甚至厌恶，进而导向算法拒绝。例如相比人类医生，人们可能对人工智能医生具有更少的情感联结或情感寄托而拒绝人工智能医生。心理联结对于决策过程的理解也很重要，尽管人类决策和算法决策都是不透明的，但由于人类之间的心理联结如感知相似性，人们认为可以通过自省的方式理解他人决策的心理过程(Nisbett, et al., 1977)，虽然这种理解并不一定是对的(Kahneman, 2003; Morewedge & Kahneman, 2010)，但这种信念使人们更偏好人类决策(Cadario et al., 2021)，而心理联结的缺失使人们无法理解算法决策的过程而拒绝算法决策。这需要更多研究，而理论框架也需要随之不断完善。

其三，强调心理动机在算法拒绝中的重要性，并不代表将心理因素置于影响算法拒绝的所有因素中的绝对主导地位。人类对算法决策的拒绝还可能出于诸如法律、政治、社会背景以及哲学思考等的各种主观原因或现实考虑。法律因素包括“数据道德”或“AI 道德”的窠

臼、人类隐私边界、个人权利与自由以及技术寡头等(彭诚信, 2020)；社会因素包括对资源分配不公平、技术性失业、阶层固化以及社会排斥等的担忧(李明倩, 2021; 孙伟平, 2021)；哲学家也对算法拒绝的原因进行了批判性的思考，认为技术化扩张可能会带来人类生存缺憾，侵蚀人类的生存意义与存在价值，使存在变得荒谬化(周露平, 2021)。虽然算法拒绝的三维动机理论强调心理因素的作用，但并非否认其他因素对算法拒绝的重要影响。算法拒绝必然是一个由多种复杂因素造成的现象，但术业专攻所致，此理论仅从心理动机来探究其原因与提高算法认同的方法。

总而言之，对算法决策的拒绝可能是由于算法主体怀疑、道德地位缺失和人类特性湮没这三维原因所致，这三种原因背后反映出信任、责任和掌控之动机。未来研究可探索算法拒绝发生的边界条件，找寻改变甚至逆转算法拒绝现象的心理变量。同时，进一步挖掘算法拒绝的其他可能心理动机，将其进行实证检验、纳入并修正理论。最后，在算法决策越来越多地以人工智能体等实体载体出现，实现工具性至社会性的转变时，宜更多地从社会性主体的角度考察人们对算法决策的态度及与之对应的心理动机。

参考文献

黄志澄. (2005). 以人为主, 人机结合, 从定性到定量的综合集成法. *西安交通大学学报: 社会科学版*, *25*(2), 55–59.

李明倩. (2021). *自动不平等：高科技如何锁定、管制和惩罚穷人*(pp. 148–171). 北京: 商务印书馆.

李月白, 江晓原. (2019). 钱学森与 20 世纪 80 年代的人工智能热. *西安交通大学学报: 社会科学版*, *39*(6), 24–29.

彭诚信(主编). (2020). *驯服算法：数字歧视与算法规制* (pp. 35–41). 上海: 上海人民出版社.

孙伟平. (2021).智能系统的“劳动”及其社会后果. *哲学研究*, (8), 30–40+128.

许丽颖, 喻丰, 邬家骅, 韩婷婷, 赵靓. (2017). 拟人化: 从 “它” 到 “他”. *心理科学进展*, *25*(11), 1942–1954.

喻丰. (2020). 论人工智能与人之为人. *人民论坛·学术前沿*, (1), 30–36.

喻丰, 韩婷婷. (2018). 有限道德客观主义的概模型. *清华大学学报(哲学社会科学版)*, *33*(157), 148–163+193.

周露平.(2021).智能拜物教的哲学性质与批判超越. *哲学研究*, (8), 41–50.

Aaltola, E. (2014). Affective empathy as core moral agency: Psychopathy, autism and reason revisited. *Philosophical Explorations*, *17*(1), 76–92.

Alserda, G. A., Dellaert, B. G., Swinkels, L., & van der Lecq, F. S. (2019). Individual pension risk preference elicitation and collective asset allocation with heterogeneity. *Journal of Banking & Finance*, *101*, 206–225.

Amiot, C. E., Sukhanova, K., & Bastian, B. (2020). Social identification with animals: Unpacking our psychological connection with other animals. *Journal of Personality and Social Psychology*, *118*(5), 991–1017.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency*, *52*(1), 7–27.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. Retrieved May 19, 2021, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Armstrong, J. S. (1980). The seer-sucker theory: The value of experts in forecasting. *Technology Review*, *82(7)*, 16–24.

Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., ... & De Souza, A. F. (2020). Self-driving cars: A survey. *Expert Systems with Applications*, 113816.

Bastian, B., Loughnan, S., Haslam, N., & Radke, H. R. (2012). Don't mind meat? The denial of mind to animals used for human consumption. *Personality and Social Psychology Bulletin*, *38*(2), 247–256.

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn. *Business & Information Systems Engineering*, *63*(1), 55–68.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34.

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences*, *23*(5), 365–368.

Biró, P., van de Klundert, J., Manlove, D., Pettersson, W., Andersson, T., Burnapp, L., Chromy, P., Delgado, P., Dworczak, P., Haase, B., Hemke, A., Johnson, R., Klimentova, X., Kuypers, D., Nanni Costa, A., Smeulders, B., Spieksma, F., Valentín, M. O., & Viana, A. (2021). Modelling and optimisation in European kidney exchange programmes. *European Journal of Operational Research*, *291*(2), 447 – 456.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576.

Bostrom, N. (2014). *Superintelligence*. Oxford: Oxford University Press.

Brewer, M. B. (1991). The social self: On being the same and different at the same time. *Personality and Social Psychology Bulletin*, *17*(5), 475–482.

Brink, K. A., Gray, K., & Wellman, H. M. (2019). Creepiness creeps in: Uncanny valley feelings are acquired in childhood. *Child Development*, *90*(4), 1202–1214.

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239.

Cadario, R., Longoni, C., & Morewedge, C. (2021). Understanding, Explaining, and Utilizing Medical Artificial Intelligence. *Nature Human Behaviour, https://doi. org/10.1038/s41562-021-01146-0*.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*(7623), 20–23.

Cinelli, M., Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118.

Chouard, T. (2016). The Go files: AI computer clinches victory against Go champion. *Nature*. http://doi.org/10.1038/nature.2016.19553

Common Cents Lab (2017, August 18). Retrieved May 19, 2021, from http://advanced-hindsight.com/commoncents-lab/

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571.

Decety, J., & Cowell, J. M. (2014). The complex relation between morality and empathy. *Trends in Cognitive Sciences*, *18*(7), 337–339.

Diehl, K., Kornish, L. J., & Lynch Jr, J. G. (2003). Smart agents: When lower search costs for quality information increase price sensitivity. *Journal of Consumer Research*, *30*(1), 56–71.

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science*, *31*(10), 1302–1314.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.

Dockrill, P. (2017). In just 4 hours, Google's AI mastered all the chess knowledge in history. *Science Alert*.

Donnelly, L. (2017), "Forget your GP, robots will 'soon be able to diagnose more accurately than almost any doctor,'" *The Telegraph*, Retrieved May 19, 2021, from https://www.telegraph.co.uk/technology/2017/03/07/robots-will-soon-able-diagnose-accurately-almost-doctor/

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021). Reducing algorithm aversion through experience. *Journal of Behavioral and Experimental Finance*, 100524.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349–379.

Fogg, B. J. (2002). *Persuasive technology: Using computers to change what we think and do*. San Francisco: Morgan Kaufmann.

Fournier, T. (2016). Will my next car be a libertarian or a utilitarian?: Who will decide?. *IEEE Technology and Society Magazine*, *35*(2), 40–45.

Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, *74*, 97–103.

Goodyear, K., Parasuraman, R., Chernyak, S., Madhavan, P., Deshpande, G., & Krueger, F. (2016). Advice taking from humans and machines: An fMRI and effective connectivity study. *Frontiers in Human Neuroscience*, *10*, 542.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619.

Gray, K., Schein, C., & Cameron, C. D. (2017). How to think about emotion and morality: Circles, not arrows. *Current Opinion in Psychology*, *17*, 41–46.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124.

Green, K. C., & Armstrong, J. S. (2007). The ombudsman: Value of expertise for forecasting decisions in conflicts. *Interfaces*, *37*(3), 287–299.

Greene, J., Rossi, F., Tasioulas, J., Venable, K., & Williams, B. (2016, March). Embedding ethical principles in collective decision support systems. In *Proceedings of 30th AAAI Conference on Artificial Intelligence* (pp. 4147-4151).

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108.

Grove, W. M., & Lloyd, M. (2006). Meehl's contribution to clinical versus statistical prediction. *Journal of Abnormal Psychology*, *115*(2), 192.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, *65*(4), 613.

Harvey, C. R., Rattray, S., Sinclair, A., & Van Hemert, O. (2017). Man vs. Machine: Comparing Discretionary and Systematic Hedge Fund Performance. *The Journal of Portfolio Management*, *43*(4), 55–69.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252–264.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, *1*(3), 333–342.

Horowitz, M. C. (2016). The ethics & morality of robotic warfare: Assessing the debate over autonomous weapons. *Daedalus*, *145*(4), 25–36.

Ireland, L. (2020). Who errs? Algorithm aversion, the source of judicial error, and public support for self-help behaviors. *Journal of Crime and Justice*, *43*(2), 174–192.

Ipsos Mori. (2020, February). Public attitudes towards online targeting - a report by Ipsos MORI for the Centre for Data Ethics and Innovation and Sciencewise (Research report). *Ipsos Mori*. Retrieved May 19, 2021, from https://www.ipsos.com/ipsos-mori/en-uk/public-attitudes-towards-online-targeting

Jago, A. S. (2019). Algorithms and authenticity. *Academy of Management Discoveries*, *5*(1), 38–56.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin, 129*(3), 339-375.

Jost, J. T., Federico, C. M., & Napier, J. L. (2009). Political ideology: Its structure, functions, and elective affinities. *Annual Review of Psychology, 60,* 307-337.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In *Proceedings of the 28th European Conference on Information Systems* (pp. 1–16).

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449-1475.

Kaya, O., Schildbach, J., AG, D. B., & Schneider, S. (2017, August 10). Robo-advice–a true innovation in asset management. *Deutsche Bank Research*. Retrieved May 19, 2021, from https://www.dbresearch.com/PROD/DBR_INTERNET_EN-PROD/ PROD0000000000449010/Robo-advice_-_a_true_innovation_in_asset_managemen.pdf.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908–911.

Komatsu, T. (2016, March). Japanese students apply same moral norms to humans and robot agents: Considering a moral HRI in terms of different cultural and academic backgrounds. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 457–458). IEEE.

Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the United States. *Humanities and Social Sciences Communications*, *8*(1), 1–11.

Kramer, M. F., Borg, J. S., Conitzer, V., & Sinnott-Armstrong, W. (2018). When do people want AI to make decisions? In *Proceedings of First Annual AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* (pp. 204–209).

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*, 633.

Kuncel, N. R. (2008). Some new (and old) suggestions for improving personnel selection. *Industrial and Organizational Psychology*, *1*(3), 343–346.

Leachman, S. A., & Merlino, G. (2017). The final frontier in cancer diagnosis. *Nature*, *542*(7639), 36–38.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270.

Lee, J. D., & Seppelt, B. D. (2006). Human factors and ergonomics in automation design. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 1570–1596). Hoboken, NJ: Wiley.

Lee, M. K., (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 2053951718756684.

Leyer, M., & Schneider, S. (2019). Me, you or AI? How do we feel about delegation. *Proceedings of the 27th European Conference on Information Systems (ECIS)*.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

Lohr, Steve (2016, October 17), "IBM is counting on its bet on Watson, and paying big money for it," *New York Times*, Retrieved May 19, 2021, from https://www.nytimes.com/2016/10/17/technology/ibm-is-counting-on-its-bet-on-watson-and-paying-big-money-for-it.html

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650.

Lourenço, C. J., Dellaert, B. G., & Donkers, B. (2020). Whose algorithm says so: The relationships between type of firm, perceptions of trust and expertise, and the acceptance of financial Robo-advice. *Journal of Interactive Marketing*, *49*, 107–124.

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*, *38*(6), 937–947.

Lusardi, A., & Mitchell, O. S. (2011). Financial literacy around the world: an overview. *Journal of Pension Economics & Finance*, *10*(4), 497-508.

Malle, B. F. (2016). Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Information Technology*, *18*(4), 243–256.

Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. In *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering, Ethics in Science, Technology and Engineering* (pp. 1–6).

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 117–124). IEEE.

McFarland, M. (2014). Elon Musk: 'With artificial intelligence we are summoning the demon.' - The Washington Post. Retrieved May 19, 2021, from https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon-musk-with-artificial-intelligence-we-are-summoning-the-demon/?utm_term=.02d648908751

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences*, *14*(10), 435-440.

Mori, M. (1970). Bukimi no tani [the uncanny valley]. *Energy*, *7*(4), 33–35.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes.

Psychological Review, *84*(3), 231.

Niszczota, P., & Kaszás, D. (2020). Robo-investment aversion. *PLoS ONE*, *15*(9), e0239277.

Oliveira, H. G. (2017, September). A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th international conference on natural language generation* (pp. 11–20).

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy.* New York: Broadway Books.

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409.

Otto, A. S., Clarkson, J. J., & Kardes, F. R. (2016). Decision sidestepping: How the motivation for closure prompts individuals to bypass decision making. *Journal of Personality and Social Psychology*, *111*(1), 1–16.

Palmeira, M., & Spassova, G. (2015). Consumer reactions to professionals who use decision aids. *European Journal of Marketing*, *49*(3/4), 302–326.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

Pew Research Center. (2019, January). Facebook algorithms and personal data (Research report). Pew Research Center. https://www.pewresearch.org/internet/2019/01/16/facebook-algorithms-and-personal-data/

Pezzo, M. V., & Pezzo, S. P. (2006). Physician evaluation after medical errors: does having a computer decision aid help or hurt in hindsight?. *Medical Decision Making*, *26*(1), 48–56.

Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted?. *Journal of Forecasting*, *36*(6), 691–702.

Promberger, M., & Baron, J. (2006). Do patients trust computers?. *Journal of Behavioral Decision Making*, *19*(5), 455–468.

Reinecke, M. G., Wilks, M., & Bloom, P. (2021). Developmental changes in perceived moral standing of robots. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43, No. 43).

Roberts, S. (2017). Christopher strachey's nineteen-fifties love machine - The New Yorker. Retrieved May 19, 2021, from https://www.newyorker.com/tech/elements/christopher-stracheys-nineteen-fifties-love-machine

Senecal, S., & Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing*, *80*(2), 159–169.

Shaffer, V. A., Probst, C. A., Merkle, E. C., Arkes, H. R., & Medow, M. A. (2013). Why do patients derogate physicians who use a computer-based diagnostic support system?. *Medical Decision Making*, *33*(1), 108–118.

Shin, K. Y., Lee, J. K., Kang, K. H., Hong, W. G., & Han, C. H. (2019). The current applications and future directions of artificial intelligence for military logistics. *Journal of Digital Contents Society*, *20*(12), 2433–2444.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... & Hassabis, D. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.

Smith, A. (2018). "Public attitudes toward computer algorithms." *Pew Research Center*. Retrieved May 19, 2021, from http://www.pewinternet.org/2018/11/16/public-attitudes-toward-computer-algorithms/

Soll, J. B., & Mannes, A. E. (2011). Judgmental aggregation strategies depend on whether the self is involved. *International Journal of Forecasting*, *27*(1), 81–102.

Starke, C., & Lünich, M. (2020). Artificial intelligence for political decision-making in the European Union: Effects on citizens' perceptions of input, throughput, and output legitimacy. *Data & Policy*, *2*, e16.

Tetlock, P. E. (2009). *Expert political judgment. How good is it? How can we know*? Princeton: Princeton University

Press.
Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. (1987). *Rediscovering the social group: A self-categorization theory*. Oxford, UK: Blackwell.
van den Berg, J., Patil, S., & Alterovitz, R. (2017). Motion planning under uncertainty using differential dynamic programming in belief space. In H. I. Christensen & O. Khatib (Eds.), *Robotics Research: The 15th International Symposium ISRR* (pp. 473–490). Cham: Springer International Publishing.
Van Swol, L. M. (2009). The effects of confidence and advisor motives on advice utilization. *Communication Research*, *36*(6), 857–873.
Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016, August). Moral judgments of human vs. robot agents. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 775–780). IEEE.
Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232.
Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117.
Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414.
Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*(2p2), 1.
Złotowski, J., Yogeeswaran, K., & Bartneck, C. (2017). Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human-Computer Studies*, *100*, 48–54.
Zuckerman, G. (2019). *The man who solved the market: how Jim Simons launched the quant revolution*. London: Penguin Random House.

A three-dimensional motivation model of algorithm aversion

ZAHNG Yuyan[1], XU Liying[2], YU Feng[1], DING Xiaojun[3], WU Jiahua[2], ZHAO Liang[4]

([1] *Department of Psychology, School of Philosophy, Wuhan University, Wuhan 430072, China*)

([2] *Department of Psychology, School of Social Sciences, Tsinghua University, Beijing 100084, China*)

([3] *Department of Philosophy, School of Humanities and Social Science, Xi'an Jiaotong University*)

([4] *Department of Publishing Science, School of Information Management, Wuhan University, Wuhan 430072, China*)

Abstract: Algorithm aversion refers to the phenomenon of people preferring human decisions but being reluctant to use superior algorithm decisions. The three-dimensional motivational model of algorithm aversion summarizes the three main reasons: the doubt of algorithm agents, the lack of moral standing, and the annihilation of human uniqueness, corresponding to the three psychological motivations, i.e., trust, responsibility, and control, respectively. Given these motivations of algorithm aversion, increasing human trust in algorithms, strengthening algorithm agents' responsibility, and exploring personalized algorithms to salient human control over algorithms should be feasible options to weaken algorithm aversion. Future research could further explore the boundary conditions and other possible motivations of algorithm aversion from a more social perspective.

Key words: algorithmic decision-making, algorithm aversion, mental motivation, human-robot interaction